DOCUMENT RESUME

ED 143 687 40 TM 006 441

AUTHOR Crandall, R. E.
TITLE A Theorem on Clustering Measures.
INSTITUTION Research Inst. for Educational Problems, Cambridge, Mass.
SPONS AGENCY Bureau of Education for the Handicapped (DHEW/OE), Washirgton, D.C.
REPORT NO RIEP-P-86
PUB DATE 73
GRANT OEG-0-8-080506-4597
NOTE 17p.; For related document, see TM 006 436

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS *Cluster Analysis; *Cluster Grouping; *Data Analysis; *Measurement Techniques; *Recall (Psychological)

ABSTRACT

In this study assumptions are made concerning the amo t of clustering in sequences containing runs of equivalent elements. It is assumed that any valid clustering measure is linear with respect to union of runs and monotone increasing under iterations of clustering operators. With these two assumptions it follows that any such measure is of necessity nonlinear in the run lengths. This requirement rules out many of the commonly used measures. A new measure is developed and discussed especially for recall studies and briefly described for applications to present experiments. (Author/MV)

****
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
****

ED143687

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

A THEOREM ON CLUSTERING MEASURES

R. E. Crandall

Research Institute for Educational Problems

29 Ware Street

Cambridge, Massachusetts 02138

PERMISSION TO REPRODUCE THIS COPYRIGHTED MATERIAL HAS BEEN GRANTED BY

MILTON BUDOFF

TO [illegible] AND ORGANIZATIONS OPERATING UNDER AGREEMENTS WITH THE NATIONAL INSTITUTE OF EDUCATION. FURTHER REPRODUCTION OUTSIDE THE ERIC SYSTEM REQUIRES PERMISSION OF THE COPYRIGHT OWNER

1973

RIEP-Print Number 86

TM006 441

Abstract

Assumptions are made concerning the amount of clustering in sequences containing runs of equivalent elements. It is assumed that any valid clustering measure is (1) linear with respect to union of runs and (2) monotone increasing under iterations of clustering operators.

With these two assumptions it follows that any such measure is of necessity nonlinear in the run lengths. This requirement rules out many of the commonly used measures. A new measure is developed and discussed especially for recall studies and briefly described for applications to present experiments.

A Theorem on Clustering Measures

Introduction

The problem of clustering in experimental psychology has motivated many investigators to find a quantitative measure of the "amount of clustering" in sequences containing several classes of equivalent objects.

In a mathematical setting, the theory of clustering goes back to the "theory of runs" [Feller, 1959] and is based on the statistical treatment of run moments [Mood, 1940] and the subsequent work of Swed and Eisenhardt [1943].

A more recent study of the theory of clustering has been applied to free recall experiments, where, in general, recall sequences exhibit categorical clustering [Bousfield (1953)]. Clustering measures have been developed specifically for recall studies, notably σ_{DB} of Dallett and Bousfield [Dallett (1964), Bousfield and Bousfield (1966)] and the related δ-score of Hudson and Dunn [1966].

In none of these derivations has a suitable definition of clustering been attempted. Indeed, this is a difficult problem. We attempt here to measure clustering in a constructive fashion by exhibiting "clustering operators" which increase, in a natural way, the amount of clustering in a sequence.

We shall then present linearity and monotonicity assumptions concerning clustering measurement. These assumptions may appear

somewhat arbitrary but in fact are implicit in most discussions of clustering. What is important here is not total validity of the assumptions but the demonstration that such assumptions immediately narrow the field of possible measures. It is in no way claimed that the measure we develop is the only true measure, rather we emphasize that departure from the specific form we derive may violate basic notions of clustering in previous theoretical treatments.

Formalism

We consider a set C of n elements together with an equivalence relation $\longleftrightarrow$ such that C is the disjoint union of t equivalence classes whose respective cardinalities are n_1, n_2, n_3, ... n_t. Here, $n = \sum_{i=1}^{t} n_i$. Denote by \$ the set of all ordered subsets of C under sequential ordering. Then any sequence $s \in \$$ admits of a natural "run" structure of the general form

$$(1) \quad s = a_1 a_2 \cdots a_R = \{ a_{j_o+1} \cdots a_{j_1} \} \{ a_{j_1+1} \cdots a_{j_2} \} \cdots$$

$$\cdots \{ a_{j_{b-1}+1} \cdots a_{j_b} \}$$

where $j_o = 0$; $j_b = R$; and

$$(2) \quad j_d < p,q \leq j_{d+1} \Rightarrow a_p \longleftrightarrow a_q$$

$$. \; j_d < p \leq j_{d+1} < q \leq j_{d+2} \Rightarrow a_p \not\longleftrightarrow a_q .$$

Thus, all elements within given brackets in (1) are equivalent by (2). We wish to represent the run structure of arbitrary s by a sequence of integers P given by

$$(3) \quad P(s) = \mu_1 \mu_2 \cdots \mu_b \quad \text{where} \quad \mu_k = j_k - j_{k-1}.$$

The sequence in (3) is just the collection of "run lengths" for the sequence s. We shall call P(s) the "cluster representation sequence" for s.

The following example should elucidate the notation. Let the set C be the collection {A,A,A,B,B,D} whence n = 6 and we shall call like letters "equivalent," so that $n_A = 3$; $n_B = 2$; $n_D = 1$. Consider the sequence s = BBCAA whose run decomposition is, from (1) s = {BB}{D}{AA}. Then P(s) is the sequence P = 212.

Given a sequence s and its "cluster representation" P(s), we further define a clustering operator K^r_{pq} on s for all pairs p, q, ∃: $a_{j_p} \longleftrightarrow a_{j_q}$ as any permutation on s satisfying:

$$(4) \quad P(K^r_{pq}(s)) = \begin{cases} \mu_1 \cdots \mu_p - \cdots \mu_q + r \cdots \mu_b & \text{if } \mu_q > \mu_p \\ \mu_1 \cdots \mu_p + r \cdots \mu_i - r \cdots \mu_b & \text{if } \mu_q \leq \mu_p \end{cases}$$

where, of course, μ_p, $\mu_q \geq r > 0$.
Those sequences which do not admit such operations will be referred to as "maximally clustered."

The Main Theorem

The following remarks should clarify the formalism inherent in (4). If a sequence s contains two equivalent but distinct runs,

a clustering operation is any permutation which transfers elements from the smaller run to the larger, but does not alter the other run lengths. If this permutation is denoted K, then P(K(s)) is obtained from P(s) by alterning the two appropriate integers. Of course, zeros in P(s) are to be suppressed; as in the following list of sequences, where each is obtained from the last by a clustering operation:

s	P(s)
ABCABA	111111
AABCBA	21111
AABBCA	2211
AAABBC	321

The last sequence here is maximally clustered, as it does not admit of any permutations satisfying (4) with its condition $r > 0$.

Clustering operations, in some sense, increase the "amount of clustering" in a given sequence. It is clear, for one thing, that any sequence must become maximally clustered after finitely many cluster operations.[2] Thus, every sequence s is the first term in a finite chain of more and more clustered sequences having the same elements. It is our hope to develop a measure which, when applied to every such chain, will yield a strictly increasing sequence of real numbers.

We now ask for such a clustering measure M which depends, for given sequence length R, only on the cluster representations (3).

Thus, M will be a function which maps finite integer sequences to the real numbers.

One of the commonly used clustering measures is essentially given by

$$(6) \quad M_{DB}(P(a_1,\ldots,a_R)) = M_{DB}(\mu_1,\ldots\mu_b) = -\frac{b}{g(R)}$$

This measure is simply related to the Dallett-Bousfield "chance clustering" Bousfield & Bousfield (1966) for a particular function g(R), by:

$$(7) \quad \sigma_{DB}(s) = \frac{t'}{g(R)} + M_{DB}(P(s)) + 1$$

where t' is the number of equivalence classes appearing in s. We find for the measure M_{DB} that, for example,

$$(8) \quad M_{DB}(711) = M_{DB}(414)$$

which requires in particular

$$(9) \quad \sigma_{DB}(AAAAAAABA) = \sigma_{DB}(AAAABAAAA).$$

From the point of view of our present discussion, the two sequences in (9) do not represent the same "amount of clustering"

in the sense that the left-hand sequence is obtained from the right-hand sequence via clustering operations of the form (4).

To develop a measure which agrees with our implicit definition of clustering, we assume that a function M on representation sequences $\{\mu_1, \ldots, \mu_b \mid b = 1, 2, \ldots,\}$ satisfies

(i) M is a linear functionwith respect to union of representation sequences

(ii) M is monotone increasing with successive application of clustering operators K.

We note here that M_{DB} of equation (6) satisfies (i) but, by (9), does not satisfy (ii). For the sake of convenience we also assume that a singleton sequence $S = a_1$ has $M(P(s)) = M(1) = 0$.

We now ask for the most general measure M satisfying the criteria (i), (ii). To this end, we prove the

Theorem: Let M be a real-valued function on sequences $\{\mu_1, \ldots \mu_b\}$ such that

(i) $M(P \cup Q) = M(P) + M(Q)$

(ii) $M(P(K(s)) > M(P(s))$ for any K, s

(iii) $M(\{1\}) = 0$.

Then

$$(10) \quad M(\mu_1, \ldots, \mu_b) = \sum_{i=1}^{b} f(\mu_i)$$

where f is a function on the positive integers with upward convexity, and f(1) = 0.†

Proof: Define, for $a_i \bullet\!\!-\!\!\bullet a_j \,|\, i, j \leq \mu$,

$$M(P(a_1 \ldots a_\mu)) = f(\mu).$$

Then, by (iii), f(1) = 0. By (i), for $b_i \bullet\!\!-\!\!\bullet b_j \not\bullet\!\!-\!\!\bullet a_1, \,|\, i,j \leq \gamma$

$$M(P(a_1 \ldots a_\mu) \cup P(b_1 \ldots b_\gamma)) = f(\mu) + f(\gamma)$$

and in general, $M(\mu_1 \ldots \mu_b) = \sum_{\mu=1}^{b} f(\mu_i)$.

Now if $P(s) = \mu_1 \ldots \mu_b$ then for $\mu_k > \mu_i$

$$M(P(K^r_{pq}(s))) = \sum_{i \neq j,k} f(\mu_i) + f(\mu_j - r) + f(\mu_k + r)$$

so that (ii) implies

$$f(\mu_k) + f(\mu_j) < f(\mu_j - r) + f(\mu_k + r)$$

so that , in particular, $f(x+1) - f(x) > f(x) - f(x-1)$

so that f is convex upward on the integers. QED

We remark that the converse of this theorem is also true, that is, if M is defined as in (10) for appropriate f, then M satisfies (i)-(iii). This follows easily from the fact that if x >y,

$$(11) \quad f(x + r) - f(x) = \sum_{j=1}^{r} f(x+j) - f(x+j-1)$$

The summation is, for convex upward f, greater than

---

† Remarks: f is upward convex here if $f(x+1)-f(x) > f(x)-f(x-1)$.

$$(12) \quad \sum_{j=1}^{r} [f(y-j+1) - f(y-j)] = f(y) - f(y-r)$$

whence $f(x+r) - f(x) > f(y) - f(y-r)$ for $x > y$ and (ii) of the theorem holds. Properties (i) and (iii) are immediate consequences of the form (10) and $f(1) = 0$.

Note that M_{DB} as given by (6) is of the form 10) but with $f(\mu)$ = constant and so the violation of monotonicity in (9) is to be expected.

A New Clustering Measure

On the basis of the aforementioned theorem and experimental work, we have developed a workable clustering measure for free recall studies.

In part of the experiment, 128 grammar school Ss were shown ($n = 18$) pictures of which 6 were animals (A), 6 were parts of the human body (B), and 6 were pieces of furniture (F). After presentation of all pictures at intervals of five seconds each, the Ss were asked to recall the objects. This procedure was repeated until each S had seen six randomizations of the 18 objects. The recall sequences were recorded and analyzed according to a commonly used measure (σ_{DB}) and the new measure, which we shall presently describe.

The theorem requires us to decide upon a convex function f. We have chosen the form $f(\mu) = \mu \log \mu$, so that

$$(13) \cdot \quad M_o(\mu_1 \cdots \mu_b) = \sum_{i=1}^{b} \mu_i \log \mu_i$$

This choice for f is computationally feasible and is, in fact, generally easier to compute than M_{DB} (equation 6).

The relevant data in the recall sequences are embodied in the numbers $\mu_1, \ldots, \mu_b$ (run lengths). Here, $R = \sum_{i=1}^{n} \mu_i$ is the total number of items recalled. For the purpose of relating clustering to recall, we used the parameter

$$(14) \qquad \bar{X} = \frac{M_o - E_R(M_o)}{\sqrt{Var_R(M_o)}}$$

where $E_R(M_o)$ is the expectation of $\sum_{i=1}^{R} \mu_i \log \mu_i$ over all possible sequences of length R constructed from the presentation list, and $Var_R(M_o)$ is the associated variance. Thus, in any particular trial, positive $\bar{X}$ corresponds to a clustering of R objects which is greater than chance given the presented list. This scheme for normalizing the clustering measure is identical with that of Hudson and Dunn [Hudson (1972)].

The computation of $E_R(M_o)$ and $Var_R(M_o)$ is extremely difficult and we have not found closed forms for these. However, a computer program was developed to ascertain these numbers for R = 1, ..., 18 using a Monte Carlo technique.

After the E_R, V_R tables were established, the σ_{DB} and $\bar{X}$ measures were plotted for the 128 subjects, and are shown in Figure 1.

Insert Figure 1 about here

The dispersion in the plot for high σ_{DB} is due to the fact that certain low recall sequences, e.g., AA, B, or FFF have maximum (1.00) σ_{DB} values. It was, in fact, this very difficulty, that σ_{DB} does not involve the presentation list, which motivated the present research.

Other aspects of this study have revealed many interesting clustering phenomena beyond the scope of this paper.

It remains to evaluate other possible measures satisfying the theorem, and, indeed, to find other definitions of clustering which may prove more meaningful.

References

Bousfield, A. K., & Bousfield, W. A. Measurement of clustering and sequential constancies in repeated free recall. Psychological Reports, 1966, 19, 935-942.

Bousfield, W. A. The occurrence of clustering in the recall of randomly arranged associates. Journal of General Psychology, 1953, 49, 229-240.

Dallett, K. M. Number of categories and category information in free recall. Journal of Experimental Psychology, 1964, 68, 1-12.

Hudson, R. L. The Hudson-Dunn clustering index revisited. Psychological Bulletin, 1972, 78, 475.

Hudson, R. L., & Dunn, J. E. A major modification of the Bousfield (1966) measure of category clustering. Behavior Research Methods and Instrumentalism, 1969, 1, 110-111.

Mood, A. M. The distribution theory of runs. Annals of Mathematical Statistics, 1940, 11, 367-392.

Swed, F. S., & Eisenhardt, C. Tables for testing randomness of grouping in a sequence of alternatives. Annals of Mathematical Statistics, 1943, 14, 66-87.

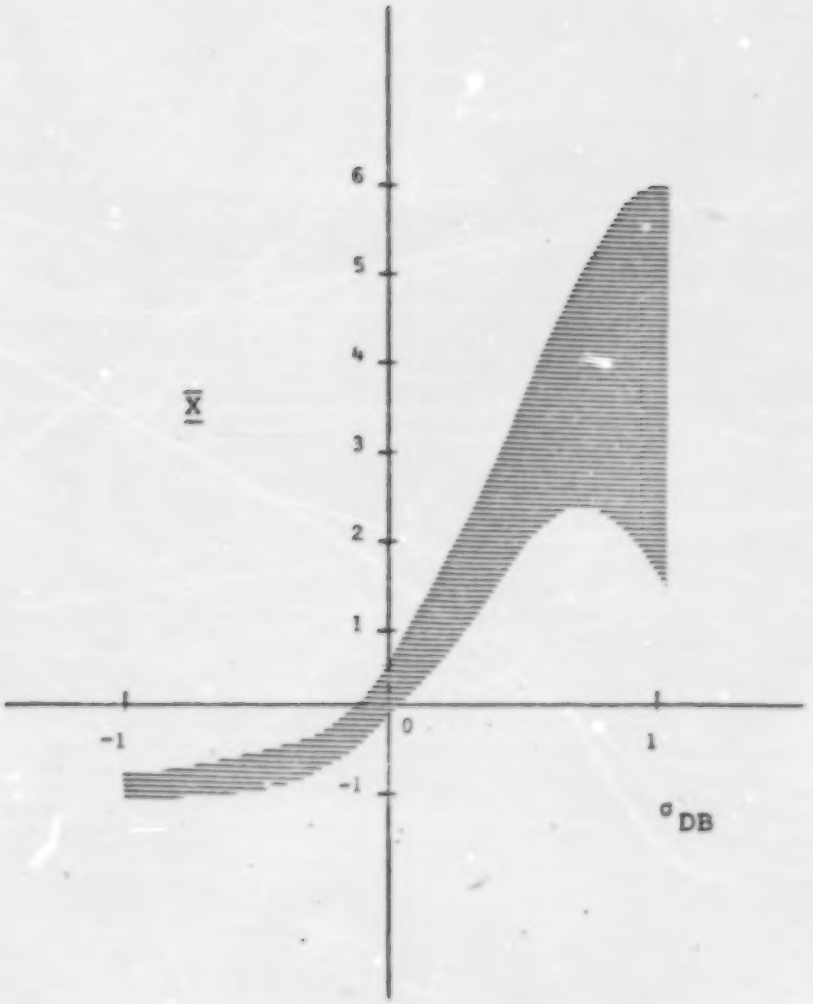
Footnotes

[1]This research was supported by Grant OEG-0-8-080506-4597 from the Bureau of Education for the Handicapped, U.S. Office of Education. The author wishes to acknowledge Joan Lerner Stein whose experimental studies of clustering in children provided the data referred to in this paper, and also Robert Harrison for his valuable suggestions concerning the form (14) for the $\underline{X}$-measure.

[2]For example, note $\Sigma\mu_i^2$ must increase after every clustering operation.

Figure 1

Experimental comparison of the nonlinear measure $\underline{\chi}$ and a standard measure q_{DB}. The number of trial points is 768 of which 95% lie in the shaded region. The dispersion for high q_{DB} is due to an anomaly for low recall (see text).


6
5
4
$\bar{X}$
3
2
1
-1
0
1
-1
σ_{DB}